Supplementary Data

DSigDB: Drug Signatures Database for Gene Set Analysis

Minjae Yoo[1,†], Jimin Shin[1,†], Jihye Kim[1], Karen A. Ryall[1], Kyubum Lee[2], Sunwon Lee[2], Minji Jeon[2], Jaewoo Kang[2] and Aik Choon Tan[1,2*]

[1]Translational Bioinformatics and Cancer Systems Biology Laboratory, Division of Medical Oncology, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA.

[2]Department of Computer Science and Engineering, Korea University, Seoul, South Korea

Contact: aikchoon.tan@ucdenver.edu

Table of Contents for Supplementary Data

DEVELOPMENT OF THE DSIGDB

Here, we describe the compilation of gene sets in Drug Signatures Database (DSigDB), a collection of drug and small molecule related gene sets based on quantitative inhibition data (See Supplementary Fig. 1). DSigDB differs from the existing resources in the following aspects: 1) DSigDB gene sets were extracted and compiled from quantitative inhibition data of drugs/compounds from a variety of databases and publications. These genes represent the direct targets of the drugs/compounds. 2) DSigDB gene sets are acquired through both automatic computational methods and manual curation. 3) Gene sets from DSigDB are explicitly designed to provide seamless integration to GSEA software (See Supplementary Fig. 2). 4) DSigDB contains the largest number of drugs/compounds related gene sets to date.

DSigDB Collections: DSigDB organizes drugs and small molecules related gene sets into four collections based on quantitative inhibition and/or drug-induced gene expression changes data (Supplementary Table 1):

D1: Approved Drugs. This collection of gene sets contains 1,202 FDA approved drugs covering 1,288 target genes. We obtained all the approved drugs from US Food and Drug Administration (FDA) website (http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm and http://fdasis.nlm.nih.gov/srs/jsp/srs/uniiListDownload.jsp). These FDA approved drug names were queried against PubChem (Wang *et al,* 2014) and ChEMBL (Bento *et al*, 2014). For each compound, we retrieved bioactivity data recorded in the BioAssays of PubChem and ChEMBL. Genes with "active" bioassay results were compiled as the drug targets. Currently, DSigDB only focus on human genes. We used Entrez Gene ID to map between databases. InChi and InChikey were used to resolve compounds ambiguity. Approval of the drugs in different regions were obtained from SWEETLEADS (Novick *et al*, 2013).

D2: Kinase inhibitors. The human kinome has been a class of intensely pursued drug targets by the pharmaceutical industry. Kinases are frequently mutated in various cancers. Therefore targeting these kinases with small molecules is an attractive therapeutic approach for personalized cancer treatment. This collection of gene sets contains 1,220 kinase inhibitors (1,065 unique kinase inhibitors) covering 407 kinases. We collected large-scale *in vitro* kinase profiling assays from literature and two databases (MRC Kinase Inhibitor database and HMS LINCS database). We considered the kinase a target of a kinase inhibitor if the IC50/Kd/Ki ≤ 1μM or the Percent of inhibition over Control (POC) ≤ 15% from the assays. These target kinases make up the gene sets for the kinase inhibitors (Supplementary Table 1).

D3: Perturbagen Signatures. This collection of gene sets was obtained from gene expression profiles induced by compounds. We collected 7,064 gene

expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02) (Lamb *et al.*, 2006). Raw microarray data were normalized by Robust Multiarray Average (RMA) within each batch using Affymetrix Power Tools. For each compound, we compared the treated vs. control gene expression profiles for each cell line. Compounds that were profiled by multiple cell lines were unified and genes that were changes with more than 2-fold change from the control were considered as gene sets were considered as the gene sets (either up and down). Compounds profiled by multiple concentrations will be regarded as different gene sets. In total, we defined 1,998 gene sets (1,154 unique compounds) covering 11,137 genes in this collection. (Supplementary Table 1).

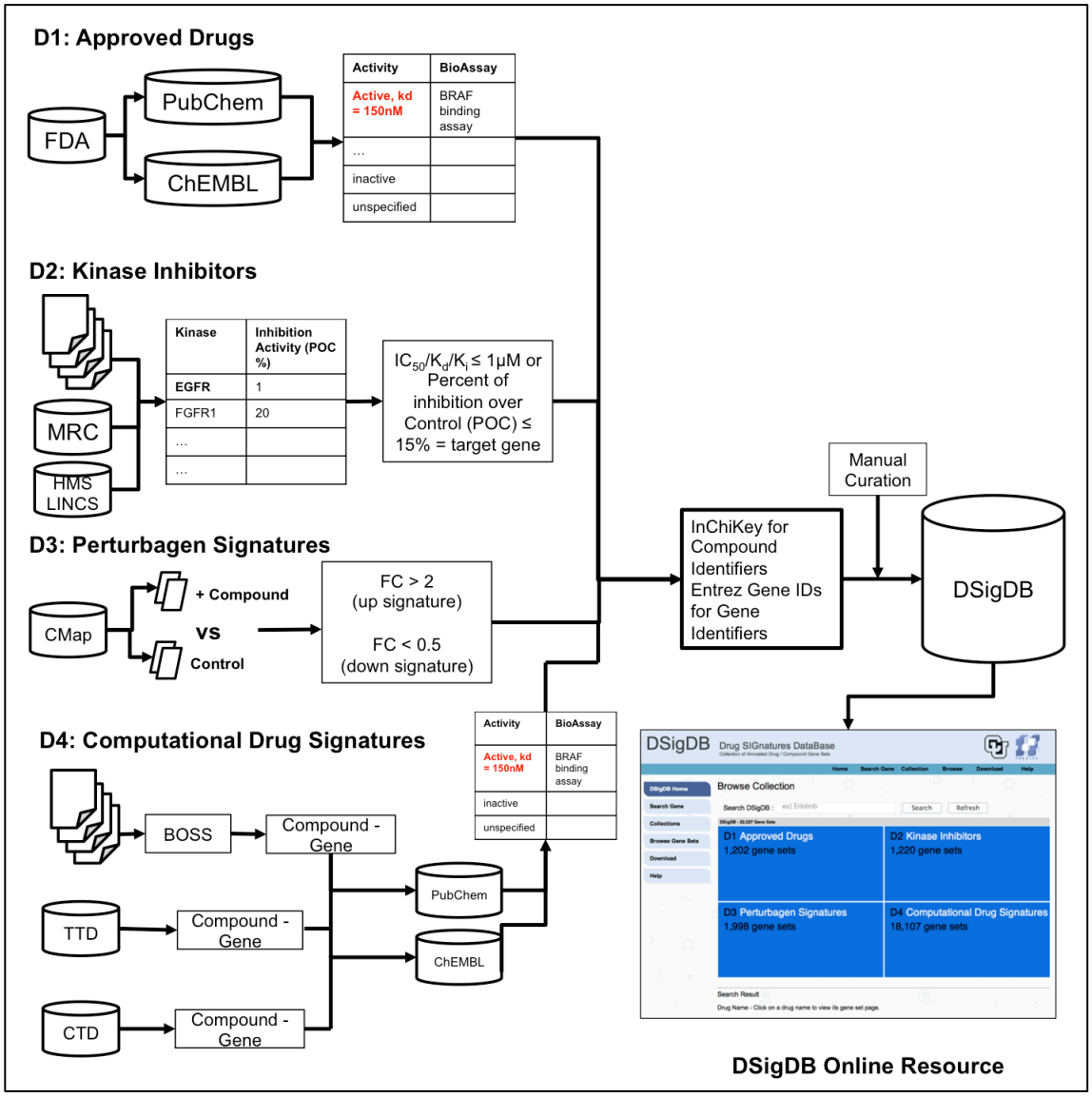
D4: Computational Drug Signatures. We compiled 18,107 drug signatures extracted from literatures using a mixture of manual curation and text mining approaches. Using manual curation of targets, we compiled 10,830 and 5,163 gene sets from the Therapeutics Targets Database (TTD) (Qin *et al.*, 2014) and the Comparative Toxicogenomics Database (CTD) (Davis *et al.*, 2013), respectively. For the text mining approach, we used the Biomedical Object Search System (BOSS) (Choi *et al.*, 2012) to acquire 2,114 co-occurrences of compounds and genes from PubMed abstracts. (Supplementary Table 1). . In addition, we also retrieved genes with "active" bioactivity data for these drugs from PubChem and ChEMBL as in D1. These genes, with quantitative inhibition data, were integrated with the drug signatures obtained from the source to construct the final gene sets for the drug

Gene set annotations: Each DSigDB gene set consists of a list of target genes of a compound. The current version of DSigDB focuses on human gene sets. We used human Entrez Gene IDs to serve as universal identifiers to map across different databases. We used InChiKey to serve as the universal compound identifiers to map between PubChem and ChEMBL, and to determine the number of unique compounds within DSigDB. As described in the DSigDB collections, these gene sets are collected from several sources and some compounds could appeared multiple times according to their source of collection. DSigDB currently holds 22,527 gene sets, consists of 17,389 unique compounds covering 19,531 genes. Statistics for the gene set size is Supplementary Table 2.

Database: After manual curation, all of the collected data were imported into a MySQL database. The DSigDB online resource retrieves and displays the data from this MySQL database. The website is implemented using Python 2.7.9, Python-CGI script for use on Unix System.

Chemical Structure: All chemical structures in DSigDB were downloaded from PubChem. The chemical compound descriptors were calculated using OpenBabel (O'Boyle *et al*, 2011). We used JSMol (JavaScript-Based Molecular Viewer From Jmol, http://www.jmol.org/) in the website to visualize the chemical structure.

File formats: DSigDB gene sets are available to download as GSEA gene set (.gmt), plain text (.txt) or detailed text (_detailed.txt) formats. The .gmt file format can be directly imported into GSEA to execute the program. The gene set results generated from GSEA provide links to the DSigDB online resource for detailed information about the compounds. The plain text format provides a simple list of gene set membership for the compound. The detailed text format provides detailed information of the relations between genes and drug. It contains four columns: Drug, Gene, Type and Source. Every line represents the relation between drug and gene, the type of interactions (either quantitative binding results or qualitative interactions), and the source of the relation (See USER MANUAL for details). We also provide these files (either .gmt, .txt or detailed.txt) for the whole database as downloadable in the Download Page.



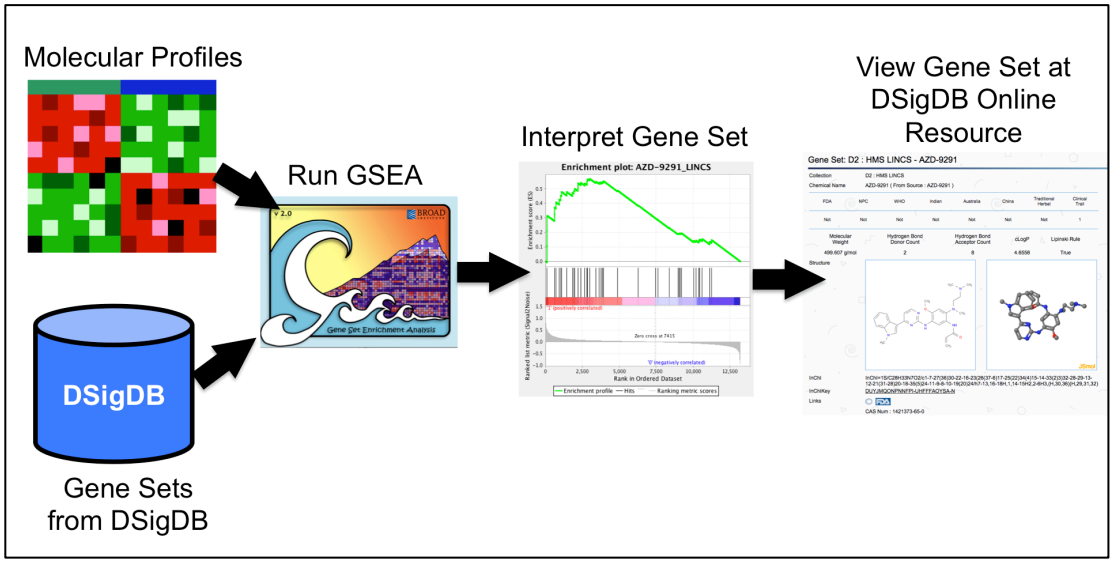
Supplementary Figure 1: DSigDB workflow.

Supplementary Table 1: DSigDB collections.

Collection (Number of gene sets)	Description	Unique Genes	Number of Gene Sets
D1: Approved Drugs (1,202 gene sets)	FDA Approved Drug	1,288	1,202
D2: Kinase Inhibitors (1,220 gene sets)	Kinase Inhibitors Gene Sets based on in vitro kinase profiling assays.	407	1,220
FDA (28 gene sets)	FDA Approved Kinase inhibitors.	341	28
HMS LINCS (90 gene sets)	Kinase inhibition assays extracted from HMS LINCS database.	381	90
MRC Kinase Inhibitor Database (157 gene sets)	Kinase inhibition assays extracted from MRC Kinome Inhibition database.	137	157
GSK (204 gene sets)	GSK Published Kinase Inhibitor Set (PKIS), kinase inhibitors used as chemical probes.	116	204
Roche (570 gene sets)	Kinase Inhibitors profiled by Roche.	153	570
KinomeScan (72 gene sets)	Kinase Inhibitors profiled by DiscoveryRx using KinomeScan technology.	374	72
RBC (99 gene sets)	Kinase Inhibitors profiled by Reaction Biology Corporation.	246	99
D3: Perturbagen Signatures (1,998 gene sets)	7,064 gene expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02).	11,137	1,998
Up signatures (985 gene sets)	Up-regulated genes when perturbed by small molecules.	8,185	985
Down signatures (1,013 gene sets)	Down-regulated genes when perturbed by small molecules.	8,642	1,013
D4: Computational Drug Signatures (18,107 gene sets)	Drug signatures extracted from literatures using a mixture of manual curation and by automatic computational approaches.	18,854	18,107
BOSS (2,114 gene sets)	Text mining approach of drug-gene targets using Biomedical Object Search System (BOSS).	3,354	2,114
TTD (10,830 gene sets)	Manual curation of targets from the Therapeutics Targets Database (TTD).	1,389	10,830
CTD (5,163 gene sets)	Curation of targets from Comparative Toxicogenomics Database (CTD).	18,700	5,163
TOTAL (22,527 gene sets)		**19,531**	**22,527**

Supplementary Table 2: Gene Set Members in DSigDB collections.

Collection	Number of Gene Sets	Number of Gene Sets with			Gene Set Members		
		≥ 5 Genes	≥ 10 Genes	≥ 15 Genes	Min	Max	Mean
D1	1,202	676	403	257	1	258	10
D2	1,220	544	334	245	1	315	15
D2 FDA	28	26	25	20	1	187	57
D2 LINCS	90	84	77	70	1	254	59
D2 MRC	157	90	53	39	1	108	13
D2 GSK	204	60	25	10	1	25	4
D2 Roche	570	166	58	26	1	80	4
D2 RBC	99	51	33	23	1	197	14
D2 Kinome	72	67	63	57	2	315	63
D3	1,998	1,253	946	796	1	3,468	81
D4	18,107	5,088	3,395	2,588	1	8,312	28
D4 BOSS	2,114	1,485	1,078	8,33	1	257	30
D4 CTD	5,163	2,748	1,738	1302	1	8,312	43
D4 TTD	10,830	1,195	492	279	1	121	3
Total	**22,527**	**7,561**	**5,078**	**3,886**	1	8,312	32



Supplementary Figure 2: DSigDB gene sets could be seamlessly integrated into GSEA software for interpreting gene sets with drugs/compounds. The DSigDB online resource provides detail molecular details for the drugs/compounds.

USE CASE EXAMPLE

To illustrate an application of DSigDB, we performed GSEA on a previously published non-small cell lung cancer (NSCLC) microarray gene expression data (Coldren *et al.*, 2006) using the D2 gene sets.

Microarray Gene Expression Profiles. All NSCLC lines were profiled by the Affymetrix HG-U133A microarrays. Raw CEL files were normalized by Robust Multiarray Average (RMA) method using Affymetrix Power Tools. The normalized gene expression profiles are available for download at http://tanlab.ucdenver.edu/DSigDB/Supplementary.

Gefitinib sensitivity. Nine gefitinib (first-generation EGFR inhibitor)-sensitive ($IC_{50} \leq 2\mu M$) and nine resistant ($IC_{50} > 4\mu M$) EGFR wild-type NSCLC lines for the analysis (Supplementary Table 3). The class label file for these NSCLC cell liens are available at http://tanlab.ucdenver.edu/DSigDB/Supplementary.

Gene Set Enrichment Analysis (GSEA). GSEA was performed comparing gefitinib sensitive vs. resistant lines using all of the D2 gene sets. We performed 1000 gene set permutations, and considered gene sets with p-value < 0.05 as significant.

Results. From the GSEA results, we observed 16 and 7 gene sets were enriched with p-value < 0.05 in the sensitive and resistant groups, respectively (Supplementary Tables 4 and 5). Notably, the top two gene sets of the sensitive group are CI-1033 and AZD9291, which are newer generation of EGFR inhibitor currently being tested in the clinic for NSCLC patients (Supplementary Table 4). According to the kinase inhibition profiles, 15 of the 16 gene sets enriched in the sensitive group inhibited EGFR (Supplementary Table 4). Conversely, none of the compounds enriched in the resistant group inhibit EGFR (Supplementary Table 5). This is expected as the comparison is between EGFR inhibitor sensitive vs resistant group. Interestingly, RO-3306, a CDK1 inhibitor was identified as enriched in resistant group (Supplementary Table 5). From the GDSC website (Yang *et al.*, 2013), two of the gefitinib-resistant lines have lower IC_{50} as compared to the four gefitinib-sensitive lines, supporting the GSEA results and suggesting that this compound may be useful for EGFR inhibitor resistant lines. (Supplementary Figure 3). All the results for this use case example are available at: http://tanlab.ucdenver.edu/DSigDB/Supplementary.

Supplementary Table 3: Gefitinib sensitivity across 18 NSCLC EGFR wild-type cell lines (Adapted from Coldren et al 2006).

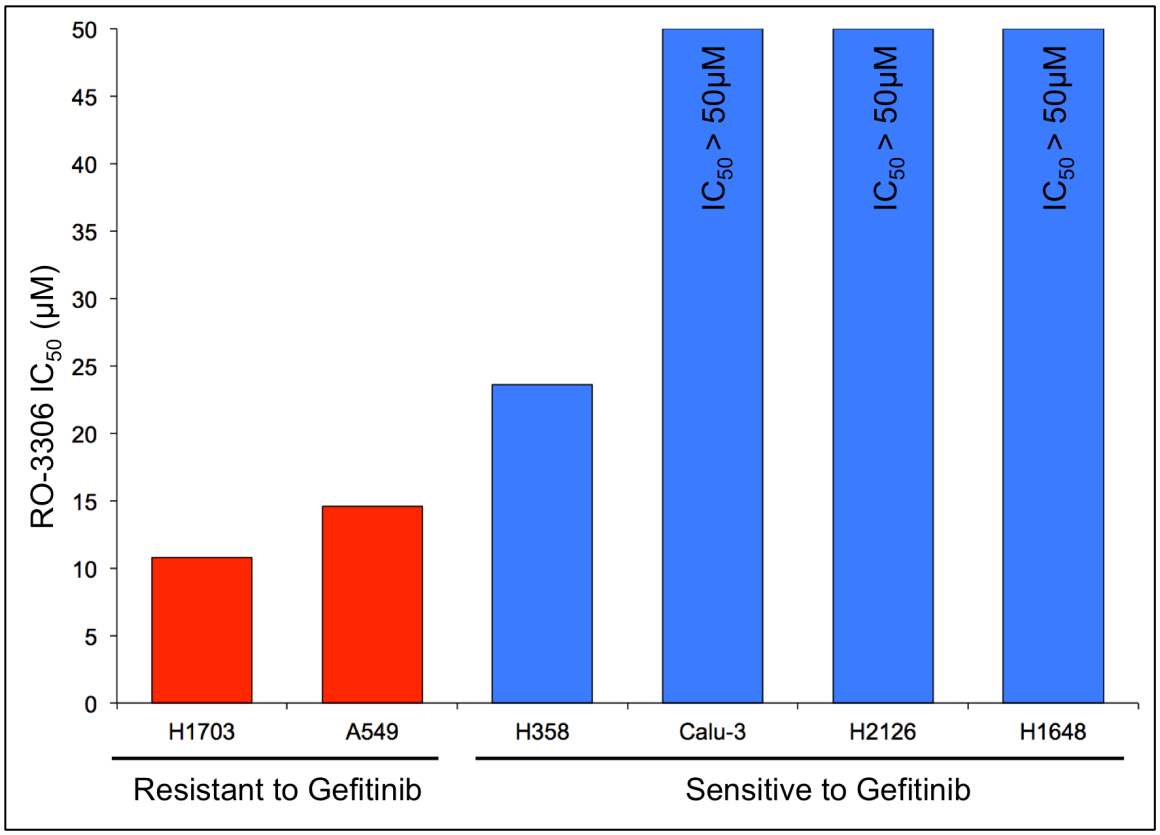
Cell line	Histology	EGFR	KRAS	Gefitinib IC50 (μmol/L)
Sensitive				
H358	BAC	Wild-type	Mutant	0.18
H322	BAC	Wild-type	Wild-type	0.25
Calu-3	Adenocarcinoma	Wild-type	Wild-type	0.3
H1334	Large	Wild-type	Wild-type	0.3
H1648	Adenocarcinoma	Wild-type	Wild-type	0.38
HCC78	Adenocarcinoma	Wild-type	Wild-type	0.4
H2126	Large	Wild-type	Wild-type	1
HCC193	Adenocarcinoma	Wild-type	Wild-type	1.5
HCC95	Adenocarcinoma	Wild-type	Wild-type	1.9
Resistant				
H125	Adenosquamous	Wild-type	Wild-type	4.8
HCC44	Adenocarcinoma	Wild-type	Mutant	7.9
H1703	Squamous	Wild-type	Wild-type	8
HCC15	Squamous	Wild-type	Wild-type	9.4
A549	Adenocarcinoma	Wild-type	Wild-type	9.6
H157	Squamous	Wild-type	Mutant	12.8
H460	Large	Wild-type	Mutant	12.9
H520	Squamous	Wild-type	Wild-type	13.6
H1299	Large	Wild-type	Wild-type	14.7

Supplementary Table 4: D2 gene sets enriched in the gefitinib-sensitive NSCLC lines sorted by Normalized Enrichment Score (p < 0.05).

GENE SET NAME	GENE SET SIZE	Normalized Enrichment Score	Nominal p-val	Intended targets	Inhibiting EGFR/ERBB2/ERBB3 based on Kinase Inhibition Assays		
					EGFR	ERBB2	ERBB3
CI-1033_KINOME SCAN	28	1.78	0.0000	EGFR/ERBB2	Yes	Yes	Yes
AZD-9291_LINCS	43	1.63	0.0125	EGFR	Yes	Yes	No
ZM-447439_LINCS	41	1.55	0.0285	AURKA	Yes	Yes	Yes
AZD-2171_KINOME SCAN	42	1.52	0.0271	VEGFR2/PDGFRA/PDGFRB	Yes	No	Yes
SB-203580_KINOME SCAN	18	1.52	0.0496	p38-alpha	Yes	No	No
WH-4-023_LINCS	124	1.48	0.0101	LCK	Yes	Yes	Yes
PP-242_KINOME SCAN	111	1.48	0.0116	MTOR/PIK3CA	Yes	Yes	No
CABOZANTINIB_FDA	45	1.47	0.0313	VEGFR2,MET	No	No	No
VANDETANIB_FDA	51	1.47	0.0355	RET/VEGFR2/EGFR	Yes	No	Yes
HG-9-91-01_LINCS	137	1.46	0.0179	SIK1	Yes	Yes	Yes
AZ-628_LINCS	51	1.46	0.0265	BRAF	Yes	No	No
VANDETANIB_KINOME SCAN	51	1.45	0.0448	RET/VEGFR2/EGFR	Yes	No	Yes
EXEL-2880/GSK-1363089_KINOME SCAN	131	1.45	0.0147	MET/AXL/VEGFR2	Yes	No	Yes
PD-173955_KINOME SCAN	105	1.40	0.0305	ABL1/SRC	Yes	No	Yes
BOSUTINIB_LINCS	69	1.40	0.0490	ABL1/SRC	Yes	Yes	Yes
R406_KINOME SCAN	183	1.39	0.0123	SYK,FLT3	Yes	No	No

Supplementary Table 5: D2 gene sets enriched in the gefitinib-resistant NSCLC lines sorted by Normalized Enrichment Score (p < 0.05).

GENE SET NAME	GENE SET SIZE	Normalized Enrichment Score	Nominal p-val	Intended targets	Inhibiting EGFR/ERBB2/ERBB3 based on Kinase Inhibition Assays		
					EGFR	ERBB2	ERBB3
KINOME_858_ROCHE	17	-1.81	0.0041	NA	No	No	No
KINOME_1901_ROCHE	17	-1.67	0.0149	NA	No	No	No
CHEMBL2062936_ROCHE	16	-1.66	0.0042	NA	No	No	No
KINOME_1242_ROCHE	16	-1.62	0.0198	NA	No	No	No
KINOME_1221_ROCHE	19	-1.62	0.0194	NA	No	No	No
KINOME_866_ROCHE	15	-1.56	0.0395	NA	No	No	No
RO-3306_MRC	29	-1.45	0.0455	CDK1	No	No	No



Supplementary Figure 3: RO-3306 sensitivity data obtained from GDSC website. Two gefitinib-resistant NSCLC cell lines have lower IC_{50} as compared to the four gefitinib-sensitive NSCLC cell lines.

REFERENCES

Bento, AP et al (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**(1),D1083-1090.

Choi,J. *et al.* (2012). BOSS: context-enhanced search for biomedical objects. *BMC Med Inform Decis Mak.*, **12**, Suppl 1:S7.

Coldren,C.D. *et al.* (2006). Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. *Mol Cancer Res.*, **4**, 521-528.

Davis,A.P. *et al.* (2013). A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database*, Nov 28:bat080.

Lamb,J. *et al.* (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929-1935.

Liberzon,A. *et al.* (2011). Molecular signature database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739-1740.

Mootha,V.K. *et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet.*, **34**, 267-273.

Novick, PA *et al* (2013). SWEETLEAD: An in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PLoS ONE*, **8**(11),e79568.

O'Boyle,NM, et al (2011). Open Babel: An open chemical toolbox. *J. Cheminf.*, **3**, 33.

Qin,C. *et al.* (2014). TTD: Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.*, **42**, D1118-D1123.

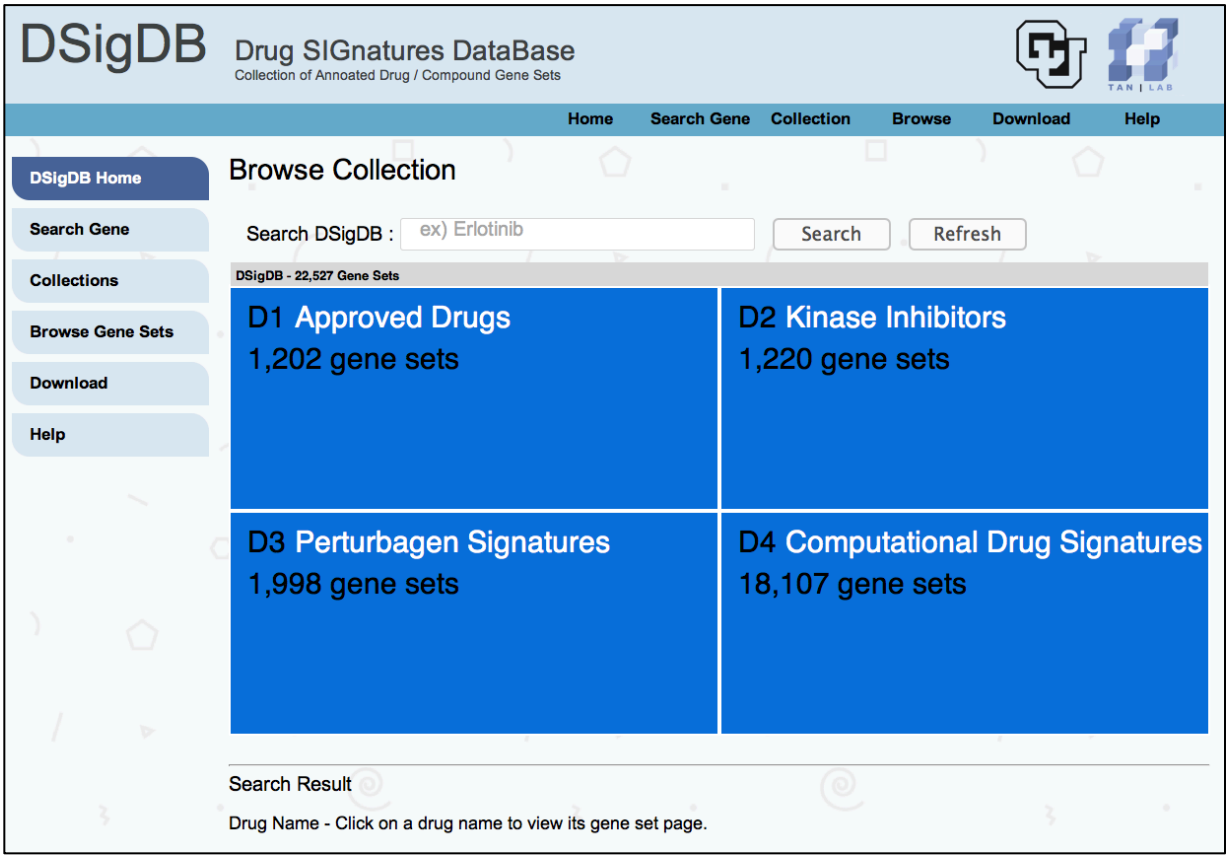
Subramaniam,A. *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.*, **102**, 15545-15550.

Wang, Y et al (2014). PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, **42**(1),D1075-82.

Yang,W. et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955-D961.

DSigDB

Drug Signatures Database
Online Resource User Manual



DSigDB Webpage: http://tanlab.ucdenver.edu/DSigDB

Version: 1.0 (May 2015)

INTRODUCTION

We report the creation of Drug Signatures Database (DSigDB), a new gene sets resource that relate drugs/compounds and their target genes, for gene set enrichment analysis. DSigDB currently holds 22,527 gene sets, representing 17,389 unique compounds covering 19,531 genes. We also develop an online DSigDB resource that allows users to search, view and download drugs/compounds and gene sets. DSigDB gene sets provide seamless integration to GSEA software for linking gene expressions with drugs/compounds for drug repurposing and translational research.

DEVELOPMENT

DSigDB is developed by the Translational Bioinformatics and Cancer Systems Biology Laboratory, Division of Medical Oncology, Department of Medicine, University of Colorado Anschutz Medical Campus.

AVAILABILITY

DSigDB is freely accessible: http://tanlab.ucdenver.edu/DSigDB.

PLEASE CITE DSigDB!

Minjae Yoo, Jimin Shin, Jihye Kim, Karen A. Ryall, Kyubum Lee, Sunwon Lee Jaewoo Kang and Aik Choon Tan. (2015). **DSigDB: Drug Signatures Database for Gene Set Analysis**. *Bioinformatics*. ***In Revision***.

CONTACT

Aik Choon Tan, aikchoon.tan@ucdenver.edu
Minjae Yoo, minjae.yoo@ucdenver.edu

ACKNOWLEDGEMENT

This work is partly supported by the Cancer League of Colorado, the National Institutes of Health P30CA046934, P50CA058187, and the David F. and Margaret T. Grohne Family Foundation. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the funders.

TABLE OF CONTENTS

1. GETTING STARTED

STARTING POINT

DSigDB (http://tanlab.ucdenver.edu/DSigDB/) is the companion online resource for search, view and download the annotated drug/compound gene sets. Figure 1 is a snapshot of the homepage of the DSigDB online resource.

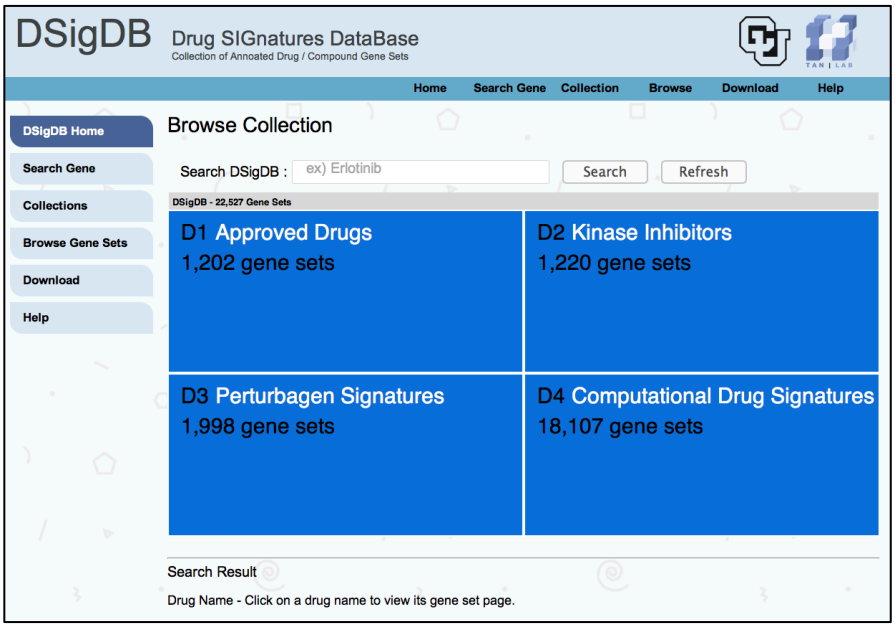


Figure 1: DSigDB Online Resource Homepage.

ANATOMY OF THE DSIGDB HOMEPAGE

Figure 2 illustrates the anatomy of the DSigDB online resource. The top and left panels represent the menu available in this website. User could search a compound/gene set by key in the name of the compound using the search box. The blue table represents the zoomable table for user to browse the DSigDB collections. The bottom section of the table represents the results page after searching or browsing.

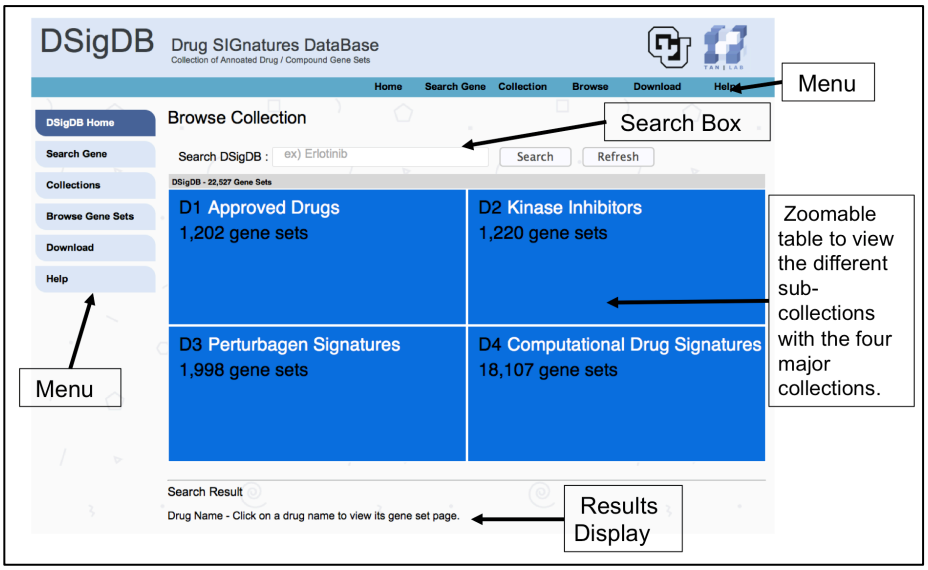


Figure 2: Anatomy of the DSigDB Online Resource Homepage.

2. SEARCHING COMPOUND IN DSigDB

To search a compound in the DSigDB, user could enter the name of the compound in the search box. For example, searching the compound “Erlotinib” (Figure 3). Once the name of the compound is entered, press the “Search” button to perform the search. The zoomable table will change from blue to red color, indicating that “Erlotinib” is found in these gene set collections. Figure 3 illustrates that “Erlotinib” is found in D1, D2 and D4 collections. At the bottom of the page, these gene sets are displayed at the results section. Click on the drug name will open a new webpage for the detail gene set in one of the collections. For a given compound query, DSigDB generates an integrated gene set from all the sources (D1 – D4) for download (.gmt and .txt files).

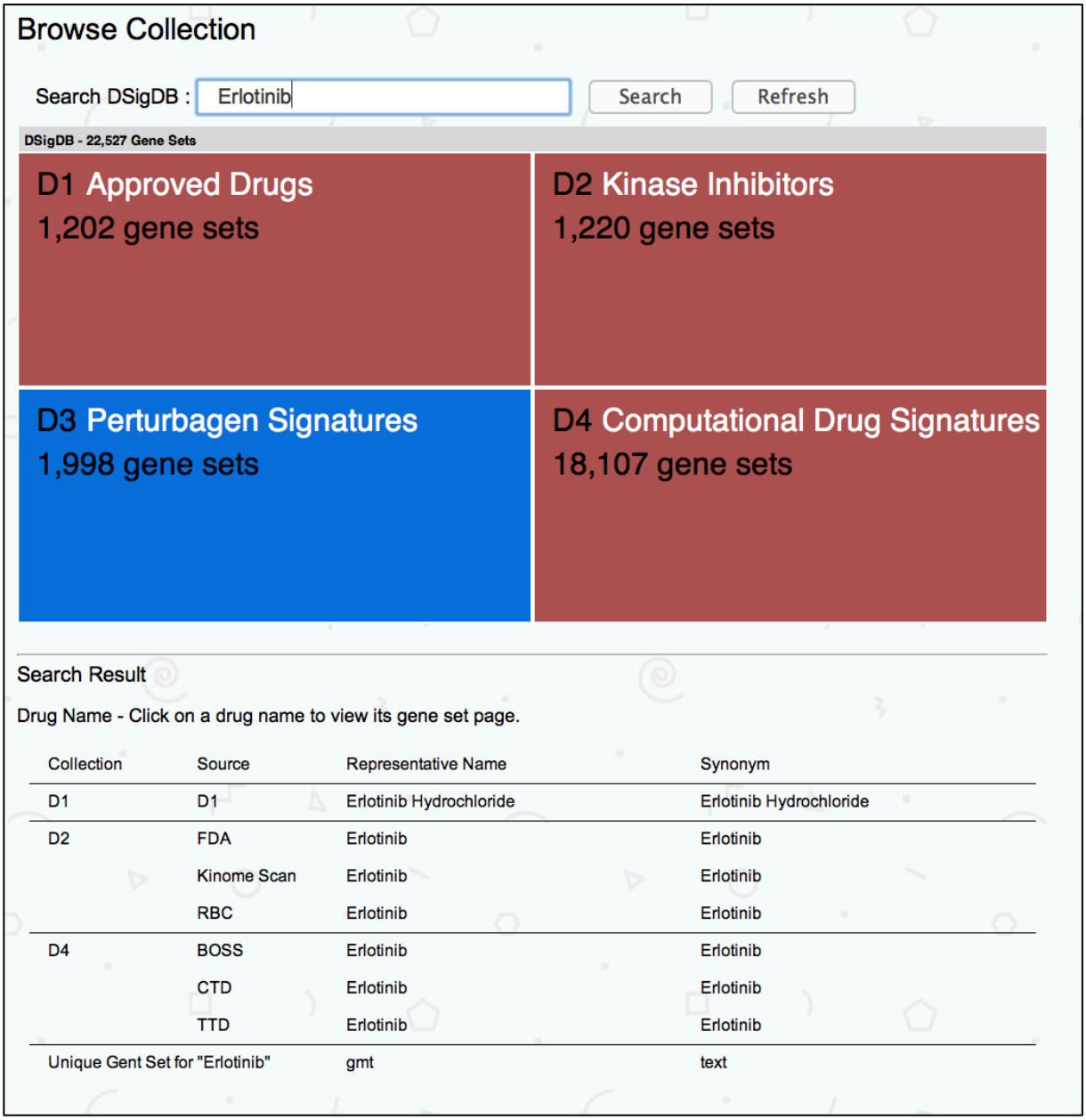


Figure 3: Searching the DSigDB.

3. SEARCHING GENE IN DSigDB

To search a gene in the DSigDB, user should click on the “Search Gene” button on the left menu (Figure 4) or on the top menu panels. The “Search Gene” page is illustrated in Figure 5.

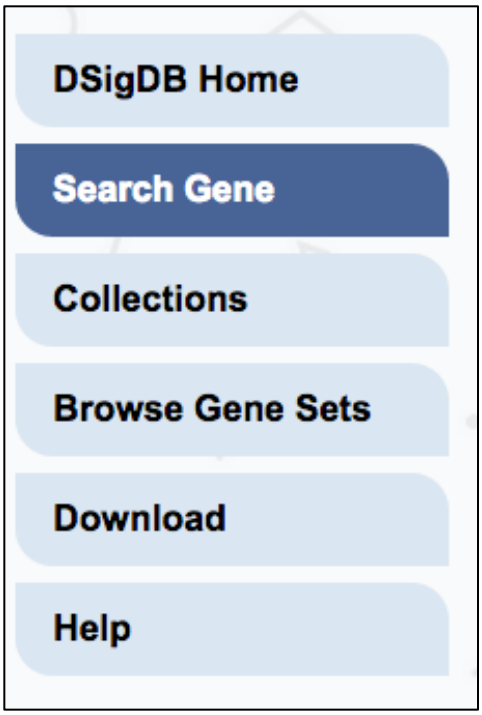


Figure 4: Search Gene option in the Left Menu.

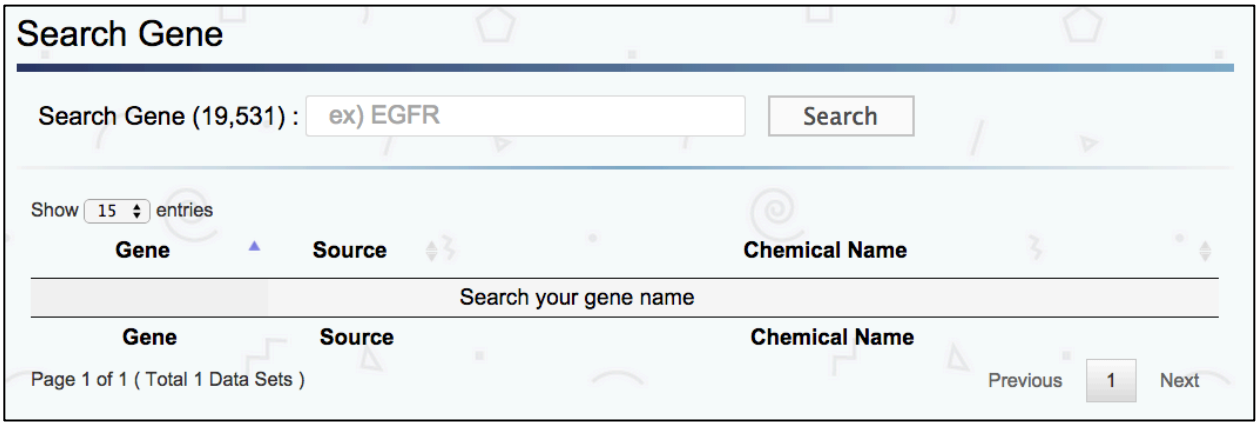


Figure 5: Screenshot of the Search Gene page.

To search for a gene that is related to a gene set in DSigDB, user could enter the official gene symbol of the gene in the search box. For example, searching the gene “EGFR” (Figure 6). Once the name of the gene is entered, press the “Search” button to perform the search. The result will refresh and display below the “Search” box. All the gene sets that contain “EGFR” as a gene member (i.e. compounds that target EGFR) will be displayed. For example, in the “EGFR” search, there are 616 gene sets that have “EGFR” as a gene member (Figure 6). Users could change the option to display the number of results per page, sort the “Source” Or “Chemical Name” by clicking the “arrow” in the results table (Figure 6).

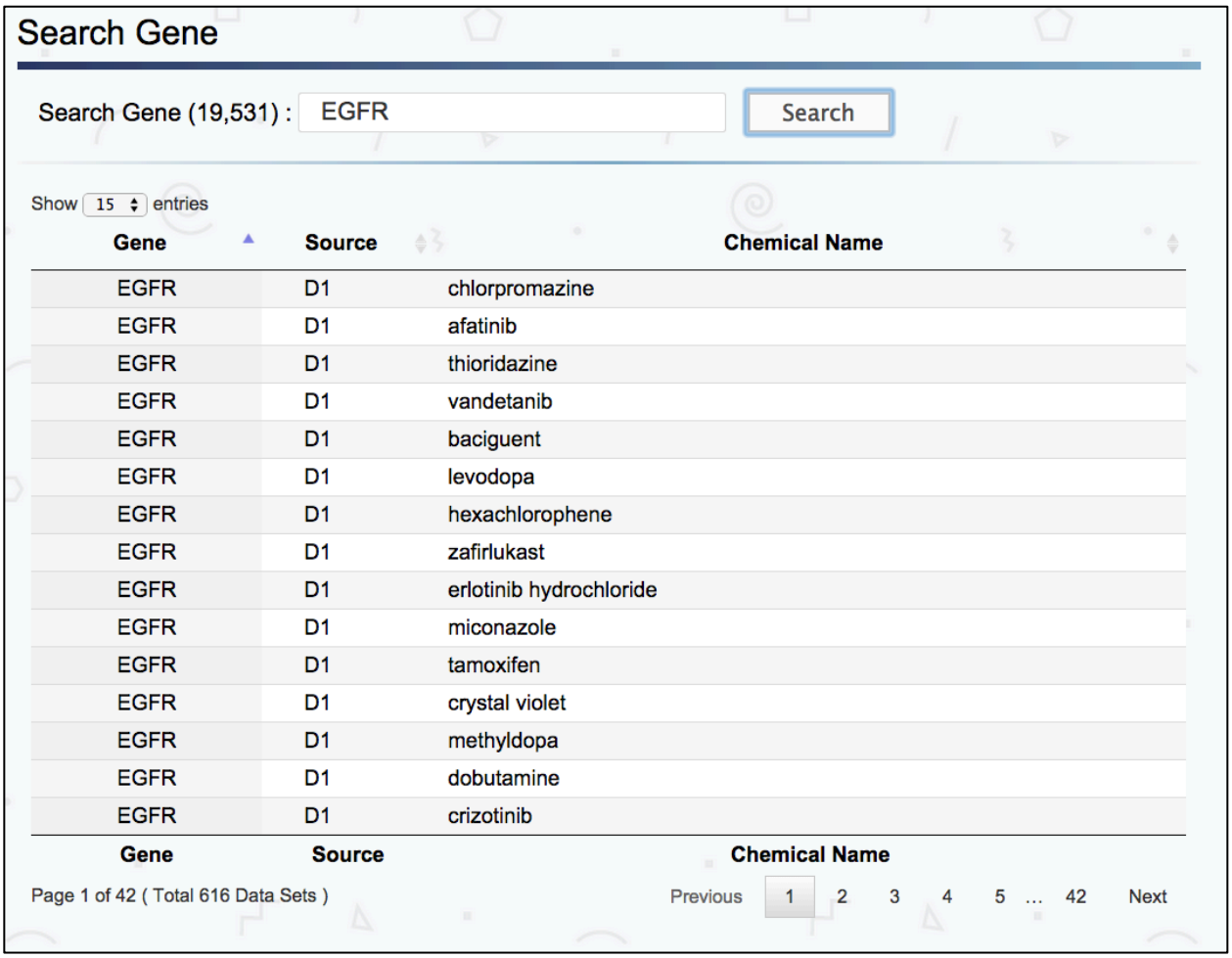


Figure 6. Search results for query “EGFR”.

4. BROWSING DSigDB COLLECTION

To browse the DSigDB collection, user may use the “Browse Collection” button on the left menu, or click on the DSigDB zoomable table (the blue square). For example, clicking on the D2: Kinase Inhibitors box (Figure 7A) will zoom in to the sub-collections of D2 (Figure 7B). There are currently seven sub-collections in the D2. To return to the original table, click on the top grey bar (Figure 7B red arrow).

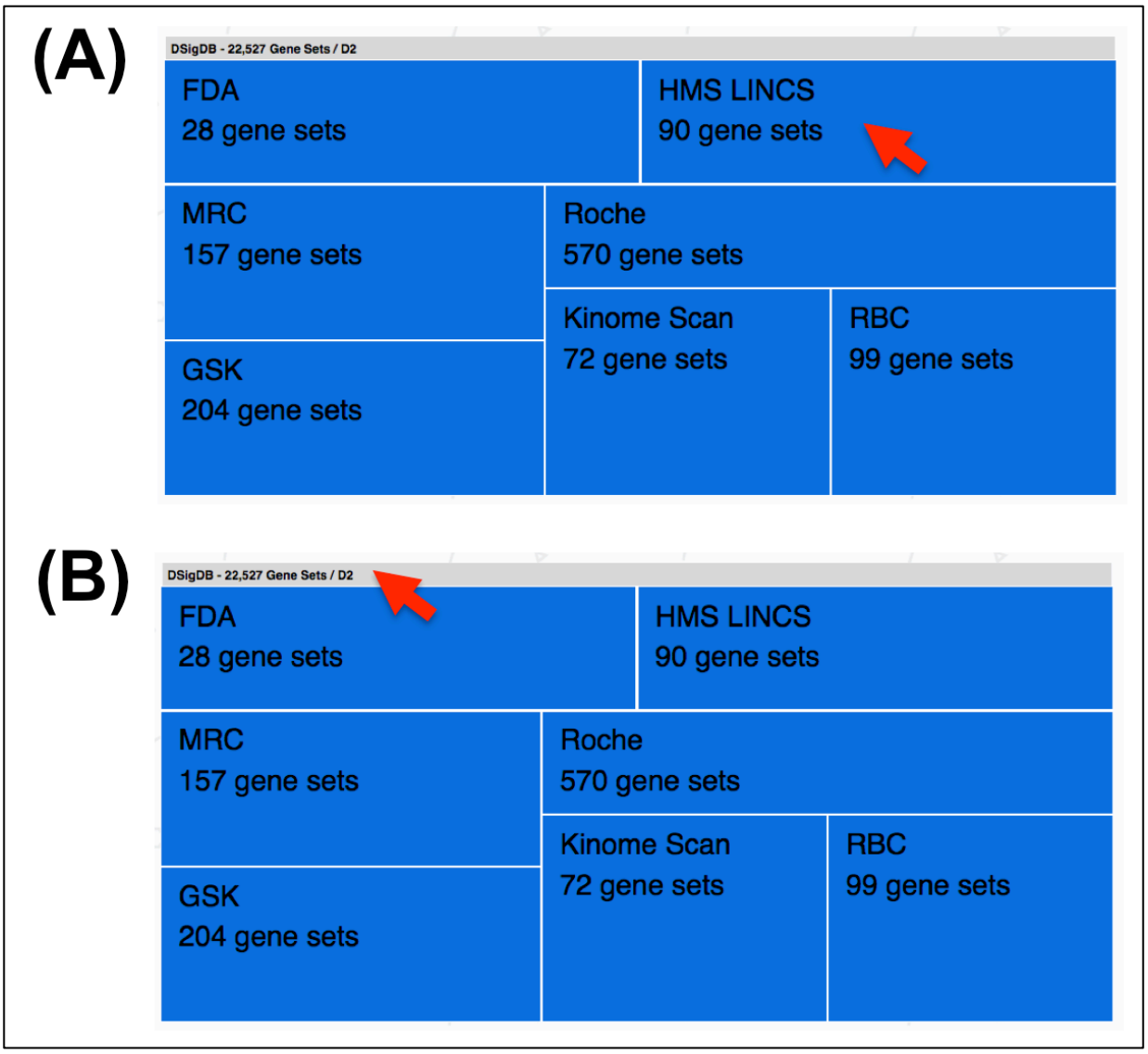


Figure 7: Browsing DSigDB using zoomable table. (A) Zooming D2: Kinase Inhibitors collection by clicking on the square. (B) There are seven sub-collections in the D2: Kinase Inhibitors. Clicking on the top grey bar will zoom out.

User could click on any of the sub-collection box. For example, by clicking the FDA (Figure 8, red arrow) of the D2: Kinase Inhibitors, the results page will list out all the FDA approved compounds that were collected in this sub-collection. Clicking on the drug will open a new window for the detail gene set page.

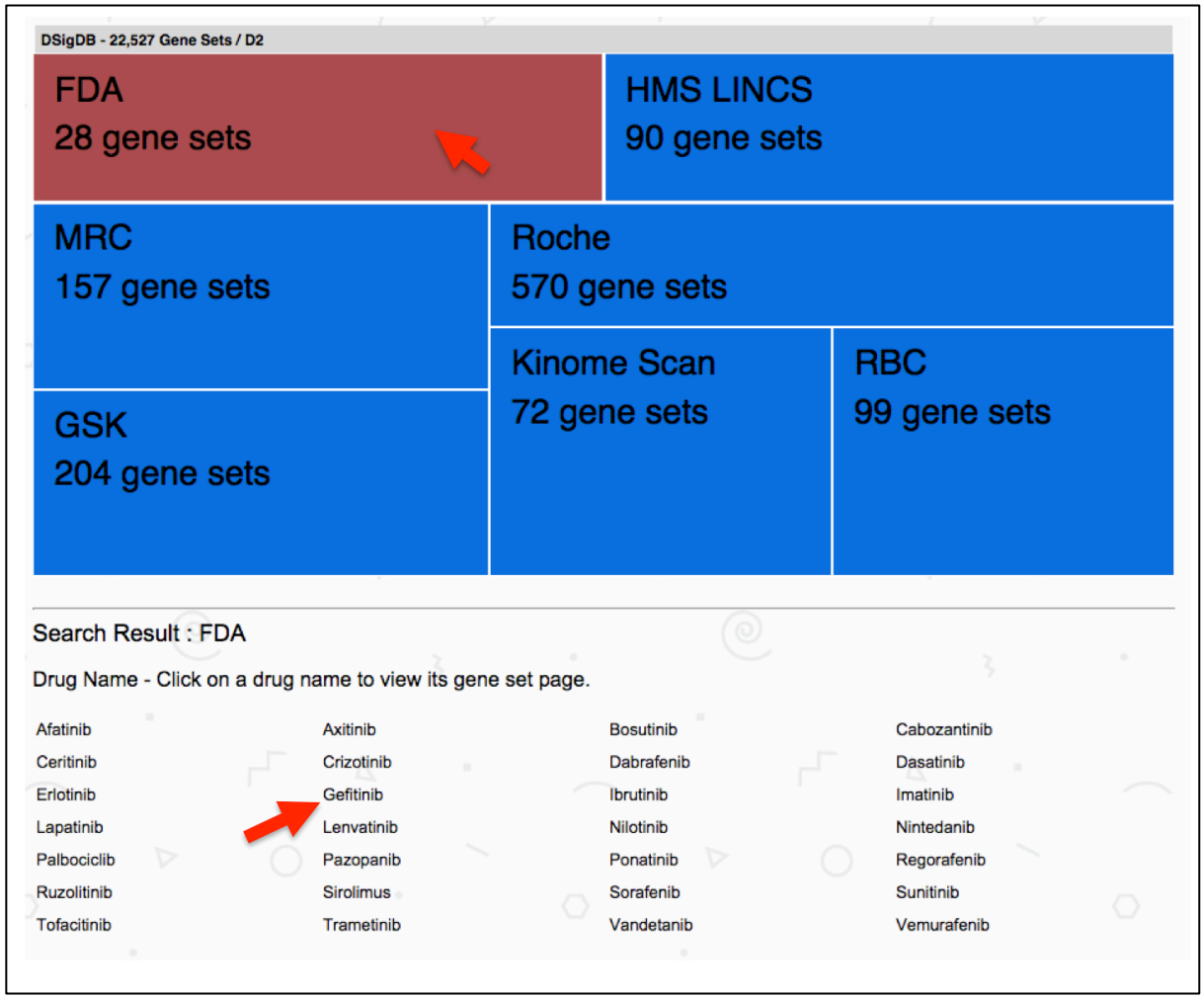


Figure 8: Browsing the FDA approved kinase inhibitors by clicking on the FDA box. At the “Results” section, it lists out the 28 kinase inhibitors and their gene sets available in DSigDB. Click on “Gefitinib” for detail view of the gene set for this drug.

5. DETAIL GENE SET WEB PAGE

Each gene set and all of its annotations are presented as an individual web page (Figure 9). Each web page contains four parts: 1) top part describes the clinical development of the compound (approved or clinical trials); 2) middle part indicates the molecular details of the compound including chemical structure (2D and 3D), links to PubChem or ChEMBL; 3) bottom part lists the gene memberships embedded links to source of evidence; 4) download of the gene set. Figure 10 illustrates the anatomy of the individual gene set page. All the external links are embedded in the web page.

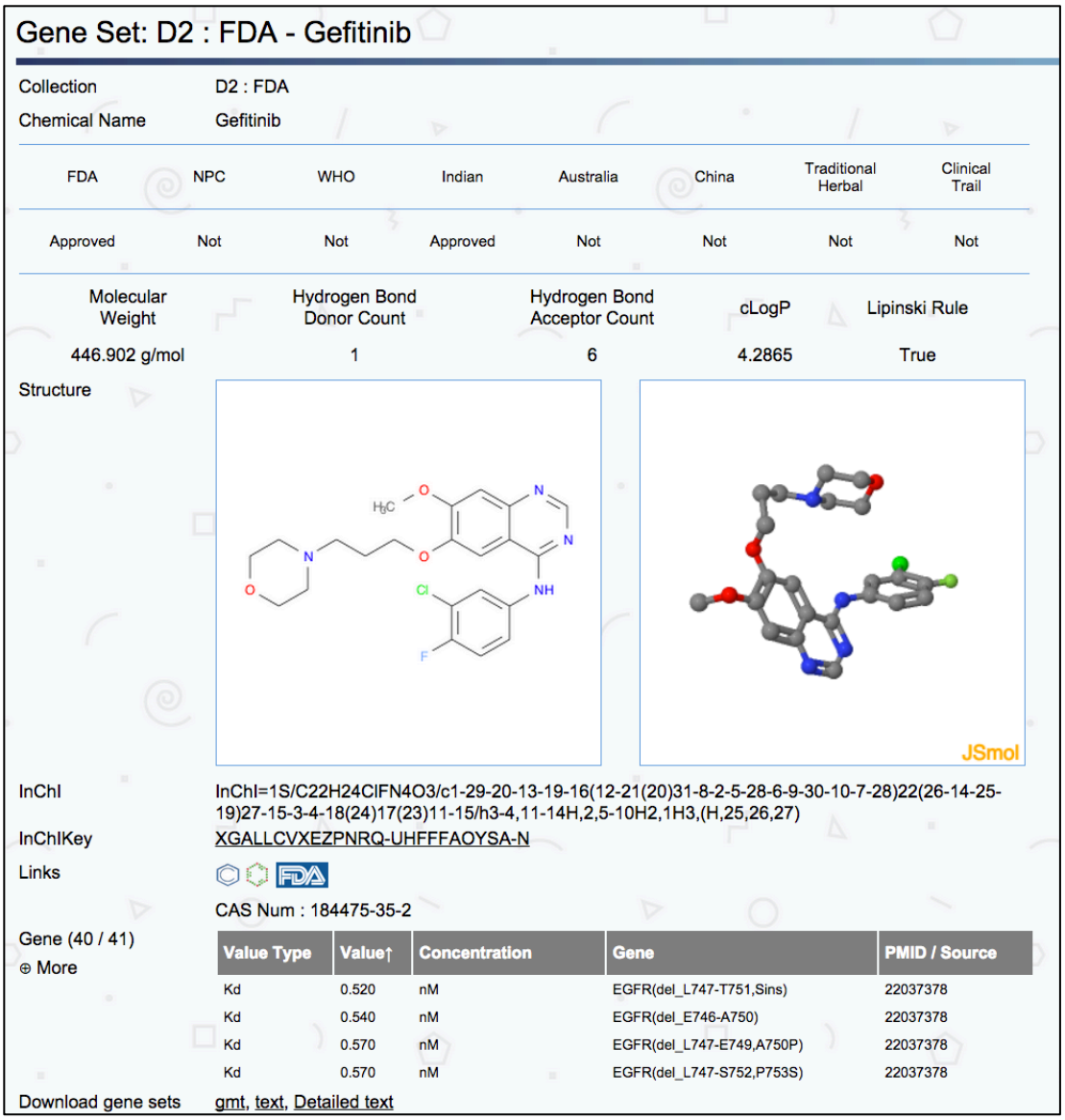


Figure 9: An example of the gene set page.

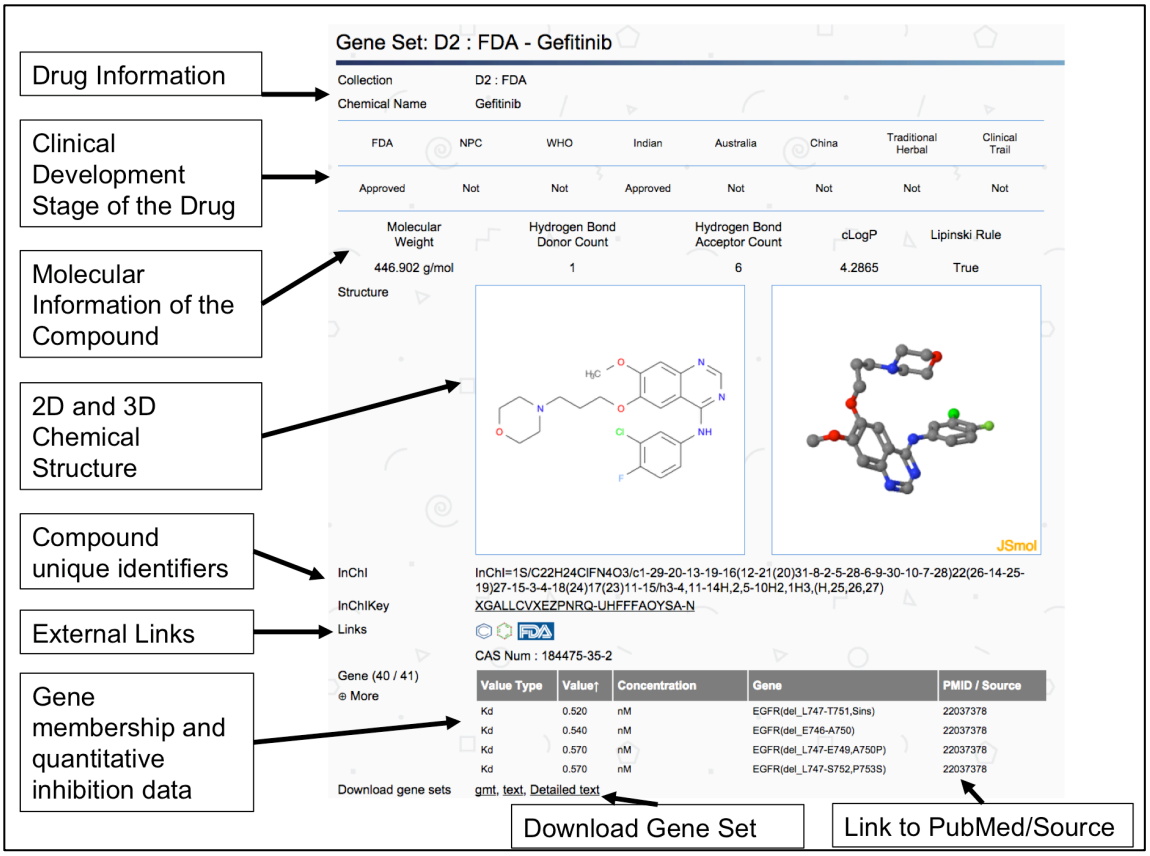


Figure 10: Anatomy of the gene set page.

DSigDB gene sets are available to download as GSEA gene set (.gmt) (Figure 11), plain text (.txt) (Figure 12) or detailed info in text (Detailed.txt)(Figure 13) formats. The .gmt file format can be directly imported into GSEA to execute the program. The gene set results generated from GSEA provides links to DSigDB online resource for detail information about the compounds.

```
Gefitinib	http://tanlab.ucdenver.edu/DSigDB/DSigDBv0.2/displayDrug.py?db=d2_fda&id=1210
	EPHA6	STK10	MKNK1	EGFR	RIPK2	MAP2K5	HIPK4	ABL1	FLT3	CSNK1E	GAK	LYN
	IRAK1	CHEK2	IRAK4	ERBB3	ERBB4	SLK	SBK1	CDK7	MAP3K19	LCK
```

Figure 11: GMT file format – Gefitinib.gmt

The plain text format provides simple listing of gene set membership of the compound. The first line contains the Compound name. The other lines represent the genes involved in this gene set. All genes are represented by their official gene symbol and separated by new line (Figure 12).

```
Compound : Gefitinib
EPHA6
STK10
MKNK1
EGFR
RIPK2
MAP2K5
HIPK4
ABL1
FLT3
CSNK1E
GAK
LYN
IRAK1
```

```
CHEK2
IRAK4
ERBB3
ERBB4
SLK
SBK1
CDK7
MAP3K19
LCK
```

Figure 12: Text file format – Gefitinib.txt

The Detailed text format provides detailed information of the relations between genes and drug. It contains four columns: Drug, Gene, Type and Source as illustrated in Figure 13. Every line represents the relation between drug and gene, the type of interactions (either quantitative binding results or qualitative), and the source of the relation.

```
Drug    Gene    Type    Source
Gefitinib       EGFR    Kd=40.0(nM)     FDA
Gefitinib       EGFR    Kd=0.54(nM)     FDA
Gefitinib       EGFR    Kd=0.98(nM)     FDA
Gefitinib       ABL1    Kd=460.0(nM)    FDA
Gefitinib       CDK7    Kd=610.0(nM)    FDA
Gefitinib       EGFR    Kd=140.0(nM)    FDA
Gefitinib       ABL1    Kd=680.0(nM)    FDA
Gefitinib       ABL1    Kd=360.0(nM)    FDA
Gefitinib       LCK     Kd=630.0(nM)    FDA
Gefitinib       ABL1    Kd=480.0(nM)    FDA
Gefitinib       MKNK1   Kd=290.0(nM)    FDA
Gefitinib       SBK1    Kd=560.0(nM)    FDA
Gefitinib       SLK     Kd=920.0(nM)    FDA
Gefitinib       EGFR    Kd=1.1(nM)      FDA
Gefitinib       ABL1    Kd=230.0(nM)    FDA
Gefitinib       IRAK4   Kd=540.0(nM)    FDA
Gefitinib       ERBB3   Kd=790.0(nM)    FDA
Gefitinib       GAK     Kd=13.0(nM)     FDA
Gefitinib       ABL1    Kd=780.0(nM)    FDA
Gefitinib       LYN     Kd=990.0(nM)    FDA
Gefitinib       IRAK1   Kd=69.0(nM)     FDA
Gefitinib       CHEK2   Kd=800.0(nM)    FDA
Gefitinib       STK10   Kd=470.0(nM)    FDA
Gefitinib       ERBB4   Kd=410.0(nM)    FDA
Gefitinib       ABL1    Kd=400.0(nM)    FDA
Gefitinib       EGFR    Kd=0.57(nM)     FDA
Gefitinib       FLT3    Kd=1000.0(nM)   FDA
Gefitinib       CSNK1E  Kd=430.0(nM)    FDA
Gefitinib       EGFR    Kd=0.52(nM)     FDA
Gefitinib       EGFR    Kd=0.94(nM)     FDA
Gefitinib       EGFR    Kd=2.0(nM)      FDA
Gefitinib       RIPK2   Kd=530.0(nM)    FDA
Gefitinib       MAP2K5  Kd=600.0(nM)    FDA
Gefitinib       ABL1    Kd=520.0(nM)    FDA
Gefitinib       EGFR    Kd=1.4(nM)      FDA
Gefitinib       HIPK4   Kd=310.0(nM)    FDA
Gefitinib       EGFR    Kd=1.0(nM)      FDA
Gefitinib       EGFR    POC=2.97(0.5uM)         FDA
Gefitinib       MAP3K19 Kd=240.0(nM)    FDA
Gefitinib       EPHA6   Kd=590.0(nM)    FDA
```

Figure 13: Detailed text file format – Gefitinib_detailed.txt

6. DSigDB COLLECTIONS

DSigDB Collections: DSigDB organized drugs and small molecules related gene sets into four collections based on quantitative inhibition data:

D1: Approved Drugs. This collection of gene sets contains 1,202 FDA approved drugs covering 1,288 target genes. We obtained all the approved drugs from US Food and Drug Administration (FDA) website, and retrieved bioactivity data for these drugs from PubChem and ChEMBL. Genes with "active" bioassay results recorded in these databases were compiled as the drug target genes

D2: Kinase inhibitors. The human kinome has been a class of intensely pursued drug targets by the pharmaceutical industry. Kinases are frequently mutated in various cancers. Therefore targeting these kinases with small molecules is an attractive therapeutic approach for personalized cancer treatment. This collection of gene sets contains 1,220 kinase inhibitors (1,065 unique kinase inhibitors) covering 407 kinases. We collected large-scale *in vitro* kinase profiling assays from literature and two databases (MRC Kinase Inhibitor database and HMS LINCS database). We considered the kinase a target of a kinase inhibitor if the $IC_{50}/K_d/K_i \leq 1\mu M$ or the Percent of inhibition over Control (POC) ≤ 15% from the assays. These target kinases make up the gene sets for the kinase inhibitors.

D3: Perturbagen Signatures. This collection of gene sets was obtained from gene expression profiles induced by compounds. We collected 7,064 gene expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02) (Lamb *et al.*, 2006). For each compound, we compared the treated vs. control gene expression profiles for each cell line. Genes with more than 2-fold change from the control were considered as gene sets (either up or down) for that compound. We defined 1,998 gene sets (1,154 unique compounds) covering 11,137 genes in this collection.

D4: Computational Drug Signatures. We compiled 18,107 drug signatures extracted from literatures using a mixture of manual curation and text mining approaches. Using manual curation of targets, we compiled 10,830 and 5,163 gene sets from the Therapeutics Targets Database (TTD) (Qin *et al.*, 2014) and the Comparative Toxicogenomics Database (CTD) (Davis *et al.*, 2013), respectively. For the text mining approach, we used the Biomedical Object Search System (BOSS) (Choi *et al.*, 2012) engine to acquire 2,114 co-occurrences of compounds and genes from PubMed abstracts. In addition, we also retrieved genes with "active" bioactivity data for these drugs from PubChem and ChEMBL as in D1. These genes, with quantitative inhibition data, were integrated with the drug signatures obtained from the source to construct the final gene sets for the drug

Gene set annotations: Each DSigDB gene set consists of a list of target genes of a compound. The current version of DSigDB focuses on human gene sets. We used human Entrez Gene IDs to serve as universal identifiers to map across different databases. We used InChiKey to serve as the universal compound identifiers to map between PubChem and ChEMBL, and to determine the number of unique compounds within DSigDB.

DSigDB Collections

DSigDB organized drugs and small molecules related gene sets into four collections based on quantitative inhibition and/or drug-induced gene expression changes data.

Collection	Description	Unique Number of Genes	Number of Gene Sets	Download
DSigDB	All Gene Sets.	19,531	22,527	GMT File
D1 : FDA Approved (browse 1,202 gene sets)	FDA Approved Drug Gene Sets.	1,288	1,202	GMT File
D2 : Kinase Inhibitors	Kinase Inhibitors Gene Sets based on in vitro kinase profiling assays.	407	1,220	GMT File
FDA (browse 28 gene sets)	FDA Approved Kinase Inhibitors.	341	28	GMT File
HMS LINCS (browse 90 gene sets)	Kinase inhibition assays extracted from HMS LINCS database.	381	90	GMT File
MRC (browse 157 gene sets)	Kinase inhibition assays extracted from MRC Kinome Inhibition database.	137	157	GMT File
GSK (browse 204 gene sets)	GSK Published Kinase Inhibitor Set (PKIS), kinase inhibitors used as chemical probes.	116	204	GMT File
Roche (browse 570 gene sets)	Kinase Inhibitors profiled by Roche.	153	570	GMT File
RBC (browse 99 gene sets)	Kinase Inhibitors profiled by Reaction Biology Corporation.	246	99	GMT File
KinomeScan (browse 72 gene sets)	Kinase Inhibitors profiled by DiscoveryRx using KinomeScan technology.	374	72	GMT File
D3 : Perturbagen Signatures (browse 1,998 gene sets)	7,064 gene expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02).	11,137	1,998	GMT File
CMAP (browse 1,998 gene sets)	7,064 gene expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02).	11,137	1,998	GMT File
D4 : Computational Drug Signatures	Drug signatures extracted from literatures using a mixture of manual curation and by automatic computational approaches.	18,854	18,107	GMT File
BOSS (browse 2,114 gene sets)	Text mining approach of drug-gene targets using Biomedical Object Search System (BOSS).	3,354	2,114	GMT File
CTD (browse 5,163 gene sets)	Curation of targets from Comparative Toxicogenomics Database (CTD).	18,700	5,163	GMT File
TTD (browse 10,830 gene sets)	Manual curation of targets from the Therapeutics Targets Database (TTD).	1,389	10,830	GMT File

Figure 14: Description of the DSigDB collections.

7. DOWNLOAD PAGE

We provide three different options to download all the data of DSigDB. Users could download the data from the Download Page. Figure 15 illustrates the screenshot of the DSigDB Download page. The page provides the version (current release is Version 1.0, May 2015), and the three file formats (.gmt, .txt and Detailed.txt) for download.

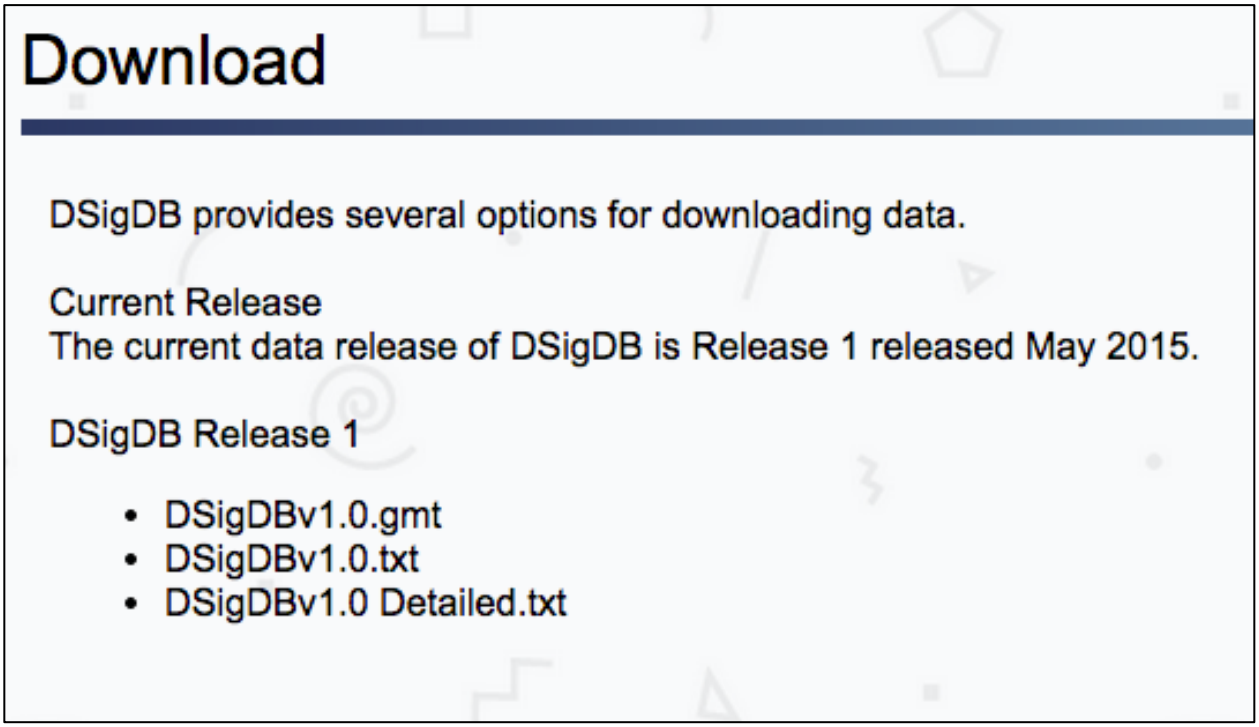


Figure 15: Screenshot of the DSigDB Download Page.

8. HELP PAGE

In the Help page, users could download a copy of this DSigDB User Manual. If users need more information, please contact:

Aik Choon Tan, aikchoon.tan@ucdenver.edu
Minjae Yoo, minjae.yoo@ucdenver.edu